

IntergenicDB: a database for intergenic sequences

Daniel Luis Notari^{1,2*}, Aurione Molin¹, Vanessa Davanzo¹, Douglas Picolotto¹, Helena Graziottin Ribeiro¹ & Scheila de Avila e Silva²

¹Centro de Computação e Tecnologia da Informação, Universidade de Caxias do Sul, Rua Francisco Getúlio Vargas, 1130 - CEP 95070-560 - Caxias do Sul, Rio Grande do Sul, Brasil; ²Instituto de Biotecnologia, Universidade de Caxias do Sul, Rua Francisco Getúlio Vargas, 1130 - CEP 95070-560 - Caxias do Sul, Rio Grande do Sul, Brasil; Daniel Luis Notari - Email: dlnotari@gmail.com; *Corresponding author

Received May 08, 2014; Revised May 23, 2014; Accepted May 24, 2014; Published June 30, 2014

Abstract:

A whole genome contains not only coding regions, but also non-coding regions. These are located between the end of a given coding region and the beginning of the following coding region. For this reason, the information about gene regulation process underlies in intergenic regions. There is no easy way to obtain intergenic regions from current available databases. IntergenicDB was developed to integrate data of intergenic regions and their gene related information from NCBI databases. The main goal of INTERGENICDB is to offer friendly database for intergenic sequences of bacterial genomes.

Availability: <http://intergenicdb.bioinfocms.com/>

Keywords: intergenic sequences, database, gene regulatory elements.

Background:

The location of coding regions as well as its regulatory sequences is an important task in theoretical analysis of a given genome [1]. Gene regulation process is essential for understanding cellular responses to environmental perturbations and virulence mechanisms of pathogens [2, 3]. The information about gene regulation process underlies in sequences as promoters, transcription factors binding sites and terminators, which are located in non-coding regions [1]. In a simplest way, a non-coding region (in bacteria) can be named as intergenic region. That means, this region comprises a part of genome located between the last nucleotide of a coding region and the first nucleotide of a subsequent coding region. The large-scale genome sequencing methods and high throughput technologies increase the genomic within the last years [4]. NCBI is considerate one of the largest and complete biological database. Apart from it, there are specific databases dedicated to store elements involved in prokaryotic gene regulation process as, RegulonDB [5], EcoGene [6], among others. The databases available have been provided input data for both motif discover

[7] and predict genomic elements [2, 8, 9, 10] approaches. However, none of these databases provide an easy-to-use way to download only intergenic regions within their biological information associated. For instance, the download of intergenic regions from NCBI is carried out by using complex queries or by developing an own computer program. The computational background is not the same for all researchers who need to run bioinformatic analysis. Hence, a specific intergenic database remains as an important lacuna. In this context, IntergenicDB (publicly database) was developed for studying intergenic sequences. This database contains a myriad of intergenic regions from 20 bacteria genomes, as well as the information of coding regions related with it.

Methodology:

IntergenicDB was developed as a free, structured and searchable source of intergenic sequences.

Dataset:

The intergenic regions of microbial genomes were downloaded from NCBI <ftp://ftp.ncbi.nih.gov/genomes/Bacteria/>.

Database design:

The data was organized by using MySQL, a relational database management system that serves as the backend for storing data.

Development:

IntergenicDB were developed using ASP.NET MVC (Model-View-Controller) website architecture, C# as programming language and IIS (Internet Information Services) as web server for hosting IntergenicDB portal. The global updates to the database take place every three months, but punctual changes are managed when necessary. These updates are intermediated by an administration user.

Database interface:

IntergenicDB general user interface is well organized and managed at the following levels: (i) the initial page presents the aims of the database; (ii) the search page allows choosing the necessary information for intergenic sequences queries, (iii) the pages with updates and publications related with intergenic sequences and (iv) others pages must be accessed using login and password. IntergenicDB has public, common and administrator user types. The administration area is restricted to the manager of the database. An overview of intergenicDB functionalities is provided in Figure 1.

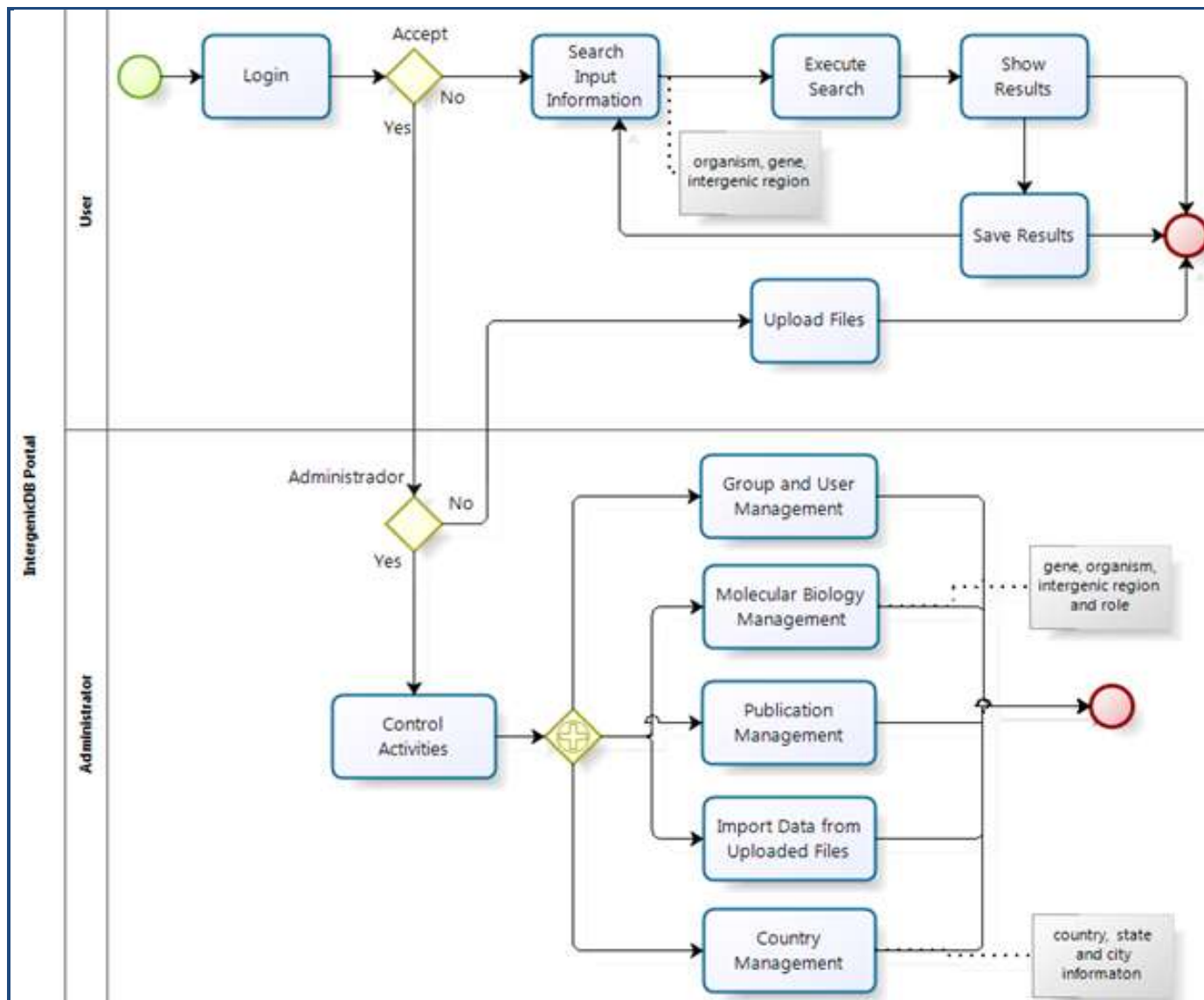


Figure 1: IntergenicDB scientific workflow.

Utility to the Biological Community:

IntergenicDB allows to users an easy way to access the information related to intergenic sequences. This forehead version of IntergenicDB supports internalization encoding to Portuguese and English languages, and it provides the search engines as follows: (i) all intergenic regions belonging a particular bacteria specie or family; (ii) an intergenic region upstream a given gene identified by its name or symbol; (iii) all intergenic regions upstream genes with specific GC nucleotide

content or a given main role; (iv) all intergenic regions with determined nucleotide length; (v) one or more intergenic regions in a specific range position at genome; (vi) all intergenic regions located in the forward or reverse DNA strand. Additionally, the user can execute queries with cross-information of the search options described above. To carry out a search, it is not necessary to complete all the fields. If one or some of they are unfilled, the returned result shows all the available data for this/these particular field/fields.

By doing a user registration, the results of the queries can be downloaded in the .txt, .xml or .csv file format. The user registration is free of charge. It is required just for statistical and database usage registration purposes. Besides the search, common users can upload intergenic sequences under supervision of database administrator. In addition to this operation, the administrator can handle activities related to user registration and management as well as database population and upgrade.

Future Developments:

As future implementations, we are committed in the improvement of the search and download areas. In this case, the user will have the option to refine even more the result provided. Another aim is the integration of the database with on line available tools which require a specific input format of intergenic sequences.

Acknowledgement:

The authors are grateful for the University of Caxias do Sul by the support for this project.

References:

- [1] Krishnamachari A *et al.* *J Theor Biol.* 2004 **227**: 429 [PMID: 15019509]
- [2] de Avila E Silva S *et al.* *J Theor Biol.* 2011 **287**: 92 [PMID: 21827769]
- [3] Janga SC & Collado-Vides J, *Res Microbiol.* 2007 **158**: 787 [PMID: 17996425]
- [4] Rangannan V & Bansal M, *J Biosci* 2007 **32**: 851 [PMID: 17914227]
- [5] Salgado H *et al.* *Nucleic Acids Res.* 2013 **41**: D203 [PMID: 23203884]
- [6] Zhou J & Rudd KE, *Nucleic Acids Res.* 2013 **41**: D613 [PMID: 23197660]
- [7] Sahota G & Stormo GD, *Bioinformatics* 2010 **26**: 2672 [PMID: 20807838]
- [8] Gardner PP *et al.* *Nucleic Acids Res.* 2011 **39**: 5845 [PMID : 21478170]
- [9] Askary A *et al.* *Genes Genet Syst.* 2009 **84**: 425 [PMID: 20228580]
- [10] Wang H & Benham CJ, *BMC Bioinformatics.* 2006 **7**: 248 [PMID: 16677393]

Edited by P Kanguane

Citation: Notari *et al.* *Bioinformation* 10(6): 381-383 (2014)

License statement: This is an open-access article, which permits unrestricted use, distribution, and reproduction in any medium, for non-commercial purposes, provided the original author and source are credited